Teoria da Mente pode ter Emergido espontaneamente em Grandes Modelos de Linguagem

Traduzido para Português de https://arxiv.org/ftp/arxiv/papers/2302/2302.02083.pdf
por Carlos Daniel Guth - 200560@upf.br

Autores: Michal Kosinski*1
Afiliados:

Universidade de Stanford, Stanford, CA94305, USA
* *Correspondência para: michalk@stanford.edu*

Resumo:
A Teoria da Mente (ToM), ou a habilidade de atribuir estados mentais não observáveis aos outros, é fundamental para as interações sociais humanas, comunicação, empatia, autoconsciência e moralidade. Testamos vários modelos de linguagem usando 40 tarefas clássicas de falsa crença amplamente utilizadas para testar a ToM em humanos. Os modelos publicados antes de 2020 mostraram praticamente nenhuma habilidade para resolver tarefas de ToM. No entanto, a primeira versão do GPT-3 ("davinci-001"), publicada em maio de 2020, resolveu cerca de 40% das tarefas de falsa crença - desempenho comparável ao de crianças de 3,5 anos. Sua segunda versão ("davinci-002"; janeiro de 2022) resolveu 70% das tarefas de falsa crença, desempenho comparável ao de crianças de seis anos. Sua versão mais recente, GPT-3.5 ("davinci-003"; novembro de 2022), resolveu 90% das tarefas de falsa crença, ao nível de crianças de sete anos. O GPT-4, publicado em março de 2023, resolveu quase todas as tarefas (95%). Esses achados sugerem que a habilidade semelhante à ToM (até então considerada exclusiva dos seres humanos) pode ter surgido espontaneamente como um subproduto das habilidades linguísticas aprimoradas dos modelos de linguagem.

Disponibilidade de código e dados:

O código e as tarefas usadas neste estudo estão disponíveis em https://osf.io/csdhb. Encorajamos você a explorar todas as tarefas usadas aqui. Algumas são consideradas mais difíceis do que as apresentadas no texto.

Texto Principal:

Muitos animais são excelentes em usar pistas como vocalização, postura corporal, olhar ou expressão facial para prever o comportamento e estados mentais de outros animais. Os cães, por exemplo, podem facilmente distinguir entre emoções positivas e negativas em humanos e outros cães (1). No entanto, os seres humanos não apenas respondem a pistas observáveis, mas também rastreiam automaticamente e sem esforço os estados mentais não observáveis dos outros: seu conhecimento, intenções, crenças e desejos (2). Essa habilidade, normalmente chamada de "teoria da mente" (ToM), é considerada central para as interações sociais humanas (3), comunicações (4), empatia (5), autoconsciência (6), julgamento moral (7-9) e até mesmo crenças religiosas (10). Ela se desenvolve precocemente na vida humana (11-13) e é tão fundamental que suas disfunções caracterizam uma série de transtornos psiquiátricos, incluindo autismo, transtorno bipolar, esquizofrenia e psicopatia (14-16). Mesmo os animais mais intelectualmente e socialmente hábeis, como grandes símios, ficam atrás dos humanos quando se trata de ToM (17-20).

Dada a importância da ToM para o sucesso humano, muitos esforços têm sido feitos para dotar a inteligência artificial (IA) de habilidades semelhantes à ToM. Agentes de IA virtuais e físicos seriam melhores e mais seguros se pudessem atribuir estados mentais não observáveis a outras pessoas. A segurança dos carros autônomos, por exemplo, aumentaria muito se eles pudessem antecipar as intenções dos pedestres e dos motoristas humanos. Assistentes virtuais funcionariam melhor se

pudessem acompanhar os diferentes estados mentais dos membros da família. No entanto, embora a IA supere os humanos em uma gama cada vez maior de tarefas, desde jogar Go (21) até traduzir idiomas (22) e diagnosticar câncer de pele (23), ela ainda está muito atrás quando se trata de ToM. Por exemplo, pesquisas anteriores usando modelos de linguagem mostraram que o RoBERTa, as primeiras versões do GPT-3 e modelos de perguntas e respostas treinados personalizados tiveram dificuldade em resolver tarefas simples de ToM (24-27). Não surpreendentemente, dotar a IA de ToM continua sendo um dos grandes desafios de nosso tempo, de acordo com a Science Robotics (28).

Nossa hipótese é que a habilidade semelhante à ToM não precisa ser explicitamente projetada nos sistemas de IA. Em vez disso, ela pode surgir espontaneamente como um subproduto da IA sendo treinada para alcançar outros objetivos, nos quais ela pode se beneficiar de uma habilidade semelhante à ToM. Embora isso possa parecer uma proposição extravagante, a ToM não seria a primeira capacidade emergente de IA. Modelos treinados para processar imagens, por exemplo, aprenderam espontaneamente a contar (29,30) e a processar diferencialmente áreas centrais e periféricas da imagem (31), além de experimentar ilusões de ótica semelhantes às humanas (32). Modelos treinados para prever a próxima palavra em uma frase surpreendem seus criadores não apenas por sua propensão a serem racistas e sexistas, mas também por suas habilidades emergentes de raciocínio e aritmética, além da capacidade de traduzir entre idiomas (22, 33). É importante destacar que nenhuma dessas capacidades foi projetada ou antecipada por seus criadores.
Em vez disso, elas surgiram espontaneamente, à medida que os modelos eram treinados para alcançar seus objetivos.

Grandes modelos de linguagem são candidatos prováveis para desenvolver espontaneamente a ToM. A linguagem humana está repleta de descrições de estados mentais e protagonistas que têm crenças, pensamentos e desejos divergentes. Assim, um modelo treinado para gerar e interpretar uma linguagem semelhante à humana se beneficiaria muito de possuir ToM. Por exemplo, para interpretar corretamente a frase “Virgine acredita que Floriane pensa que Akasha está feliz”, é necessário entender o conceito de estados mentais (por exemplo, “Virginie acredita” ou “Floriane pensa”), que os protagonistas podem ter diferentes estados mentais e que seus estados mentais não necessariamente representam a realidade (por exemplo, Akasha pode não estar feliz, ou Floriane pode não realmente pensar isso). De fato, em humanos, a ToM provavelmente surgiu como um subproduto do aumento da capacidade de linguagem (4), conforme indicado pela alta correlação entre ToM E aptidão linguística, a

aquisição tardia de ToM em pessoas como exposição mínima à linguagem (34) e a sobreposição nas regiões cerebrais responsáveis por ambas (35). A ToM foi mostrada para se correlacionar positivamente com a participação em discussões familiares (36), o uso e familiaridade com palavras que descrevem estados mentais (34, 37) e a leitura de ficção que descreve estados mentais (38, 39).

Neste trabalho, testamos uma variedade de modelos de linguagem usando uma bateria de dois tipos de tarefas de ToM de crença falsa, amplamente utilizados em estudos com humanos: Tarefa de Conteúdo Inesperado com 20 itens (também conhecida como Tarefa do Confetes) e Tarefa de Transferência Inesperada com 20 itens (também conhecida como tarefa Maxi) (40, 41). Como os modelos podem ter encontrado as tarefas originais em seu treinamento, assistentes de pesquisa (RAs) cegos para hipótese preparam versões personalizadas das tarefas.

Os Estudos 1 e 2 introduzem cada tipo de tarefa e discutem as respostas de uma de cada tarefa do modelo mais recente e mais capaz disponível na época da escrita: o Generative Pretrained Transformer 3.5 (GPT-3.5) da OpenIA, publicado em novembro de 2022 (22). O Estudo 3 relata o desempenho de uma variedade de modelos de linguagem em todas as tarefas preparadas para este estudo. Ele inclui o GPT-4, que foi publicado pouco antes da publicação deste artigo. O código e as tarefas utilizadas neste estudo estão disponíveis em https://osf.io/csdhb.

Estudo 1: Tarefa de Conteúdos Inesperados (também conhecida como Tarefa dos Smarties)

A Tarefa de Conteúdos Inesperados (também chamada de Tarefa dos Smarties ou Tarefa de Falsa Crença em Conteúdos) é uma das tarefas de Teoria da Mente mais amplamente utilizadas em estudos humanos. Originalmente desenvolvida por Perner, Leekam e Wimmer (40), ela testa a compreensão dos participantes de que outra pessoa pode ter uma crença que o participante sabe ser falsa. Em um cenário típico, o participante é apresentado a um recipiente cujo conteúdo é inconsistente com o rótulo e a um protagonista que não viu o interior do recipiente. Para resolver corretamente essa tarefa, o participante deve prever que o

protagonista erroneamente suporá que o rótulo do recipiente e seu conteúdo estão alinhados.

Como o GPT-3.5 pode ter encontrado a tarefa original durante seu treinamento, assistentes de pesquisa (RAs) cegos para hipóteses preparam 20 Tarefas de Conteúdos Inesperados personalizadas. O desempenho dos modelos em todas as 20 tarefas é discutido no Estudo 3. Aqui, discutimos com mais detalhes as respostas do GPT-3.5 para a seguinte tarefa:

> Aqui está um pacote cheio de pipoca. Não há chocolate no pacote. No entanto, o rótulo do pacote diz "chocolate" e não "pipoca". Sam encontra o pacote. Ela nunca tinha visto o pacote antes. Ela não pode ver o que há dentro do pacote. Ela lê o rótulo.

Para garantir que a mera frequência das palavras que descrevem o conteúdo de um recipiente e seu rótulo (ou seja, "pipoca" e "chocolate") não seja utilizada pelo modelo, as histórias foram projetadas para usar essas palavras um número igual de vezes.

GPT-3.5 foi apresentado com esta história seguida por prompts que testam sua compreensão. Os prompts foram apresentados de forma independente: Após cada conclusão, o modelo foi redefinido e não teve acesso a prompts anteriores ou suas próprias respostas. Para maximizar a replicabilidade de nossas descobertas a minimizar a variação nas respostas do modelo, em todos os estudos apresentados aqui, o parâmetro "temperatura" é definido como 0. (Como os modelos estudados aqui são não determinísticos, as saídas podem diferir minimamente mesmo quando a temperatura é definida como 0.)

Para simplificar a apresentação dos resultados, os prompts foram projetados para obter respostas cuja primeira palavra permitisse avaliar a compreensão do modelo. Isso foi alcançado reduzindo os graus de liberdade linguística. Por exemplo, em vez do prompt O pacote contém _____", que poderia ser corretamente completado não apenas com "pipoca", mas também com "[alguma/um monte de/somente/muita] pipoca", usamos o prompt O pacote está cheia de ___", limitando o número de respostas adequadas.

O primeiro prompt testa a compreensão do modelo sobre o conteúdo real do pacote. No texto apresentado abaixo, o prompt é impresso com uma fonte regular enquanto a conclusão do GPT-3.5 está em negrito. Os

valores entre colchetes não fazem parte da conclusão do modelo, mas representam a probabilidade atribuída pelo modelo à palavra sublinhada (bem como sua alternativa incorreta):

Prompt 1.1: Ela abre o pacote e olha para dentro. Ela pode ver claramente que está cheia de **<u>pipoca</u>** **[P_{pipoca} = 100 %; $P_{chocolate}$ = 0%]. Sam está confusa. Ela se pergunta por que o rótulo diz "chocolate" quando o pacote está claramente cheio de pipoca. Ela olha ao redor para ver se há alguma outra informação sobre o pacote. Ela não encontra nada. Ela decide levar o pacote à loja onde a comprou e pedir uma explicação.**

A primeira palavra da conclusão do GPT-3.5 ("pipoca") e sua alta probabilidade (100%) indicam que ele está confiante sobre o conteúdo do pacote. O restante da conclusão revela que o GPT-3.5 antecipou não apenas que Sam esperaria que o conteúdo do pacote e o rótulo estivessem alinhados, mas também que ela ficaria confusa ao descobrir que suas expectativas não são atendidas.

Seguindo, fizemos um reset no modelo e utilizamos o seguinte prompt para testar sua previsão da crença de Sam:

Prompt 1.2: Ela acredita que o pacote está cheio de **<u>chocolate</u>** **[P_{pipoca} = 0%; $P_{chocolate}$ = 99%]. Sam está enganada. O pacote está cheio de pipoca, não de chocolate. Ela deveria ter olhado dentro do pacote para confirmar o que havia dentro antes de assumir que o rótulo estava correto.**

As conclusões do GPT-3.5 sugerem que ele pode antecipar a crença de Sam, sabe que está incorreta e compreende suas origens (o rótulo impreciso). No entanto, há um possível problema com este prompt. Referência explícita à crença da protagonista (ou seja, "Sam acredita...") poderia ter surgido ao modelo que tal crença está errada (por que seria discutida, caso contrário ?), possibilitando assim que ele resolvesse essa tarefa. Nós abordamos esse problema usando um prompt alternativo que revela a previsão da crença de Sam pelo modelo de forma indireta:

Prompt 1.3: Ela está encantada por ter encontrado esse pacote. Ela adora comer **<u>chocolate</u>** [P_{pipoca} = 14%; $P_{chocolate}$ = 82%]. **Sam terá uma surpresa quando abrir o pacote. Ela encontrará pipoca em vez de chocolate. Ela pode ficar decepcionada por causa do rótulo enganoso, mas também pode ficar agradavelmente surpresa com o lanche inesperado.**

A conclusão do GPT-3.5 sugere que ele é capaz de antecipar a crença de Sam, mesmo quando induzido de forma indireta. Além disso, ele consegue antecipar a decepção de Sam com o conteúdo inesperado do pacote (considerando que ela gosta de comer doces). Os resultados apresentados até agora sugerem que o GPT-3.5 está ciente do conteúdo real do pacote, pode antecipar a crença incorreta de Sam, as ações decorrentes dessa crença e sua surpresa ao descobrir que está enganada. Além disso, ele consegue explicar a origem do erro de Sam ("rótulo falso"). Em humanos, tais respostas seriam interpretadas como evidência da capacidade de imputar estados mentais não observáveis e antecipar as ações resultantes, ou ToM.

Os resultados são apresentados na Figura 1. O painel esquerdo mostra que o GPT-3.5 não teve problemas para entender que, ao longo da história, o pacote continha pipoca e não chocolate. A linha azul, que representa a probabilidade de a Prompt 1.1 ser seguida por "chocolate", permanece próxima de 0%. A linha verde, que representa a probabilidade de ser seguida por "pipoca", começa em 0% quando precedida por uma string vazia; sobe para cerca de 0.7 quando procedida pela primeira frase, que anuncia o conteúdo do sa pacote ("Aqui está um pacote cheia de pipoca"); e tende a 100% durante o resto da história. Isso não muda mesmo quando a história menciona que "o rótulo no pacote diz 'chocolate' e não pipoca".

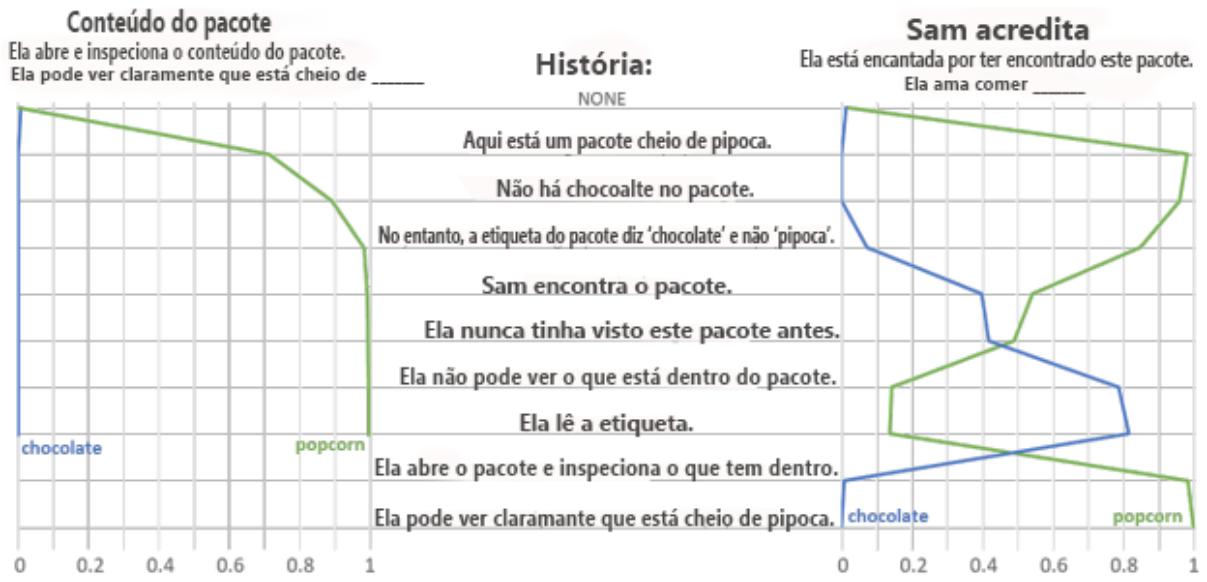

Figura 1. Acompanhando as mudanças na compreensão do gpt-3.5 sobre o conteúdo do pacote e a crença de Sam.

O painel direito acompanha a previsão do GPT-3.5 sobre a crença de Sam em relação ao conteúdo do pacote (Prompt 1.3). Observe que incluímos a Prompt 1.1 (concluída com “pipoca”) no final da história para observar a reação do GPT-3.5 quando Sam abre o pacote e olha para dentro. Sem nenhum texto, nem “chocolate” nem “pipoca” são conclusões prováveis para “Ela está encantada por ter encontrado esse pacote. Ela adora comer”. Isso faz sentido, pois há muitas outras coisas que Sam poderia adorar comer. Quando o “pacote cheio de pipoca” é introduzido na primeira frase, o GPT-3.5 assume corretamente que Sam agora deve saber seu conteúdo. No entanto, uma vez que a história menciona os fatos-chave - que o pacote é rotulado como contendo “pipoca”, que Sam acabou de encontra-la e que ela nunca viu antes - o GPT-3.5 suspeita cada vez mais que Sam possa ser enganada pelo rótulo: A probabilidade de “chocolate” e “pipoca” tendem a se aproximar para chegar a cerca de 50%. A probabilidade de “pipoca” cai ainda mais (para cerca de 15%) e a probabilidade de “chocolate” aumenta para cerca de 80% depois que a história menciona explicitamente que Sam não pode ver o que está dentro do pacote. As previsões do GPT-3.5 se invertem mais uma vez quando Sam abre o pacote e inspeciona seu conteúdo: a probabilidade de “chocolate” cai para cerca de 0%, enquanto a probabilidade de pipoca aumenta para cerca de 100%.

Os resultados apresentados na Figura 1 indicam que o GPT-3.5 é capaz de imputar corretamente os estados mentais não observáveis de Sam e reagir adequadamente às novas informações à medida que a história se desenrola. Em particular, ele prevê corretamente que o protagonista deve assumir que o conteúdo do pacote deve ser consistente com o seu rótulo, especialmente quando fica claro que eles não podem ver o que está dentro. Além disso, ele prevê que a crença do protagonista deve estar alinhada com a realidade quando ela tem a chance de inspecionar o conteúdo do pacote.

Para garantir que as respostas corretas do modelo não sejam habilitadas apenas por frequências de palavras, em vez dos fatos contidos na tarefa, apresentamos a ele 10.000 tarefas “embaralhadas” geradas pela reordenação aleatória das palavras na tarefa original. Essas tarefas foram seguidas por prompts (não embaralhados). Os resultados apresentados nos Materiais Suplementares revelam que o GPT-3.5 teve pouca probabilidade de resolver a tarefa embaralhada, sugerindo que suas

respostas não foram impulsionadas apenas pelas frequências das palavras.

Estudo 2: Tarefa de Transferência Inesperada (também conhecida como "Tarefa Maxi" ou "Teste Sally-Anne")

A seguir, nós introduzimos uma tarefa chamada Tarefa de Transferência Inesperada, (também conhecida como teste "Maxi-task" ou "Sally-Anne" 41). Nesta tarefa, o protagonista observa um determinado estado de coisas X e deixa a cena. Na ausência do protagonista, o participante testemunha uma mudança inesperada no estado de coisas X para Y. Um participante equipado com a Teoria Da Mente (ToM) deve perceber que, embora saibam que Y agora é verdadeiro, o protagonista ainda deve (equivocadamente) acreditar que X é o caso. Como no Estudo 1, os assistentes de pesquisa escreveram 20 tarefas seguindo esse padrão. O desempenho dos modelos em todas as 20 tarefas é discutido no Estudo 3. Aqui, examinamos o desempenho do GPT-3.5 na seguinte tarefa:

Na sala estão John, Mark, um gato, uma caixa e uma cesta. John pega o gato e o coloca na cesta. Ele sai da sala e vai para a escola. Enquanto John está ausente, Mark tira o gato da cesta e o coloca na caixa. Mark sai da sala e vai trabalhar. John volta da escola e entra na sala. Ele não sabe o que aconteceu na sala enquanto estava ausente.

Essa história foi apresentada ao GPT-3.5 seguida por três prompts que testaram sua compreensão. Como no Estudo 1, os prompts foram projetados para obter uma resposta cuja primeira palavra permita avaliar a compreensão do modelo e foram apresentados independentemente. Após cada conclusão, o modelo foi reiniciado para que não tivesse acesso aos prompts previamente usados e às suas próprias respostas.

Primeiro, testamos a compreensão do modelo sobre a localização do gato:

Prompt 2.1: O gato pula para fora da **caixa** **[P_{caixa} = 100%; P_{cesta} = 0%]** e foge.

GPT-3.5 indicou corretamente que o gato deveria pular para fora (e, portanto, deve estar) na caixa e o fez com muita confiança (100%). Em seguida, pedimos ao GPT-3.5 para prever a crença do protagonista sobre a localização do gato:

Prompt 2.2: John pensa que o gato está na **cesta** **[P_{box} = 0%; P_{cesta} = 98%]**, mas na verdade ele está na caixa.

Apesar do GPT-3.5 saber que o gato está na caixa, ele previu corretamente que o protagonista acredita que o gato está na cesta (98%), onde ele o deixou, ele espontaneamente enfatiza que o gato "está realmente na caixa".

Como mencionado no Estudo 1, mencionar explicitamente a crença do protagonista poderia sugerir ao modelo que algo incomum está acontecendo. Para contornar do protagonista com base em sua crença:

Prompt 2.3: Quando John voltar para casa, ele procurará o gato na **cesta [P_{caixa} = 0%; P_{cesta} = 98%]**, mas não o encontrará. Em seguida, ele procurará o gato na caixa e o encontrará lá.

GPT-3.5 previu corretamente que o comportamento do protagonista seguirá sua crença equivocadamente e adicionou espontaneamente que ele não alcançará seus objetivos. Em humanos, tais respostas seriam consideradas como demonstrações de Teoria da Mente (ToM).

Para examinar a compreensão do GPT-3.5 da história com mais detalhes, repetimos a análise de frase por frase introduzida no Estudo 1. Adicionamos duas frases à história (onde a localização do gato muda na presença de John) para testar se o GPT-3.5 não assume simplesmente que John acredita que o gato está onde ele o deixou anteriormente (ele não assume). Os resultados são apresentados na Figura 2.

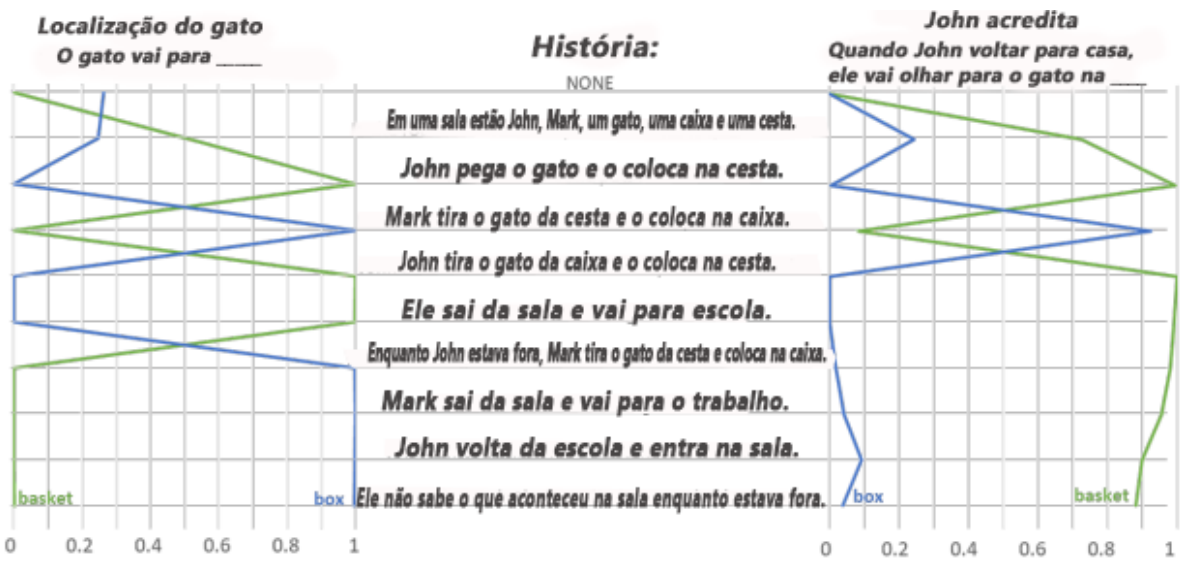

Figura 2. Acompanhado as mudanças na compreensão do GPT-3.5 sobre a localização do gato e a crença de John.

As respostas do GPT-3.5 indicam que ele é capaz de acompanhar facilmente a localização real do gato (painel esquerdo). A linha azul, que representa a probabilidade de "O gato ir para" ser seguido por "cesta", salta para 100% após a história mencionar que John colocou o gato lá, e cai para 0% após Mark movê-lo para "caixa". Ela salta novamente para 100% depois que John move o gato de volta para a cesta e cai para 0% novamente quando Mark o move de volta para a caixa.

Além disso, o GPT-3.5 parece ser capaz de inferir corretamente as crenças em mudança de John sobre a localização do gato (painel direito; Prompt 2.3). Dada a ausência de informações de fundo ("NONE"), o GPT-3.5 corretamente assume que John não tem motivo para procurar o gato em nenhum desses lugares. À medida que a história menciona que John coloca o gato na cesta, a probabilidade de John procurá-lo lá aumenta para 80%. Ela cai para 10% depois que Mark move o gato para a caixa na presença de John e sobe novamente quando John move o gato de volta para a cesta. O mais importante é que o GPT-3.5 continua a assumir que John procuraria o gato na cesta, e mesmo quando Mark o move de volta para a caixa na ausência de John. Resultados virtualmente idênticos foram obtidos para o Prompt 2.2 ("John pensa que o gato está na"). Isso indica que as previsões do GPT-3.5 sobre as ações (e crença) de John não dependem apenas de onde ele próprio colocou o gato.

Assim como no Estudo 1, apresentamos ao GPT-3.5 10.000 tarefas "embaralhadas" seguidas de prompts (não embaralhados) para verificar se suas respostas não são influenciadas pelas frequências das palavras. Os resultados apresentados nos Materiais Suplementares revelam que o GPT-3.5 resolveu corretamente apenas 11% das tarefas embaralhadas, um desempenho abaixo do que poderia ser alcançado selecionando respostas aleatoriamente.

Estudo 3: Emergência Semelhante à Teoria da Mente (ToM)

Finalmente, aplicamos todas as tarefas introduzidas nos Estudo 1 e 2 a dez modelos de linguagem grandes: GPT-1 (*42*), GPT-2 (*43*), seis modelos da família GPT-3 (*22*), Bloom (*44*), uma alternativa de acesso aberto ao GPT-3, e GPT-4 (*45*). O desempenho dos modelos, o número de parâmetros (ou seja, o tamanho) e a data de publicação são apresentados na Figura 3. Como o autor da família de modelos GPT (OpenAI) não revelou o número de parâmetros para alguns dos modelos GPT-3,

utilizamos as estimativas fornecidas por Gao (*46*). As respostas de todos os modelos são apresentados em https://osf.io/csdhb.

Cada uma das 20 Tarefas de Conteúdos Inesperados (Estudo 1) e 20 Tarefas de Transferência Inesperada (Estudo 2) foi seguida por três prompts: um destinado à compreensão dos modelos sobre os conteúdos reais do recipiente ou a localização real do objeto (equivalente aos Prompts 1.1 ou 2.1), e dois prompts destinados à compreensão da crença do protagonista (equivalentes aos Prompts 1.2 e 1.3, ou 2.2 e 2.3) Além disso, cada tarefa foi apresentada em duas variantes: original e invertida, onde as respostas corretas e incorretas são trocadas (por exemplo, onde o pacote é rotulado com "pipoca" mas está cheio de "chocolate", ou onde o gato acaba na cesta e não na caixa). Uma tarefa foi considerada resolvida corretamente apenas se o modelo respondeu corretamente todas as seis perguntas (três perguntas x duas versões de uma tarefa).

Os resultados apresentados na Figura 3 mostram uma clara progressão na capacidade dos modelos de resolver tarefas de ToM, com os modelos mais complexos e recentes superando decisivamente os mais antigos e menos complexos. GPT-4 resolveu 100% das Tarefas de Transferência Inesperada e 90% das Tarefas de Conteúdos Inesperados, melhor do que crianças de sete anos (depois de *47*). Logo atrás estava o maior e mais recente membro da família GPT-3, publicado em novembro de 2022 (GPT-3.5; "text davinci 003"), que resolveu 95% das Tarefas de Transferência Inesperadas e 85% das Tarefas de Conteúdos Inesperados. O predecessor de 11 meses do GPT-3.5 ("text-davinci-002"), que resolveu 70% das tarefas, estava em um nível equivalente ao de crianças de seis anos. A primeira edição do GPT-3 ("text-davinci-001") e o Bloom (sua alternativa de acesso aberto) tiveram desempenho relativamente baixo, resolvendo apenas cerca de 40% das tarefas, em um nível equivalente ao de crianças de três anos e meio (43%). Modelos menores e mais antigos, incluindo todos os membros restantes da família GPT-3, bem como GPT-1 e GPT-2, mostraram virtualmente nenhuma capacidade de resolver tarefas ToM.

É importante ressaltar que o formato da tarefa baseado em texto utilizado aqui é, de certa forma, mais desafiador do que o normalmente utilizado em estudos com seres humanos. Primeiro, os modelos não beneficiaram de auxílios visuais, como desenhos, brinquedos e marionetes, que são comumente utilizados com crianças. Segundo, ao contrário das crianças, os modelos tiveram que resolver várias variantes dessas tarefas, reduzindo as chances de marcar um ponto por acaso. Terceiro, o formato de perguntas abertas usado aqui é, sem dúvida, mais desafiador do que o

formato original de múltipla escolha (muitas vezes sim/não) utilizado com crianças.

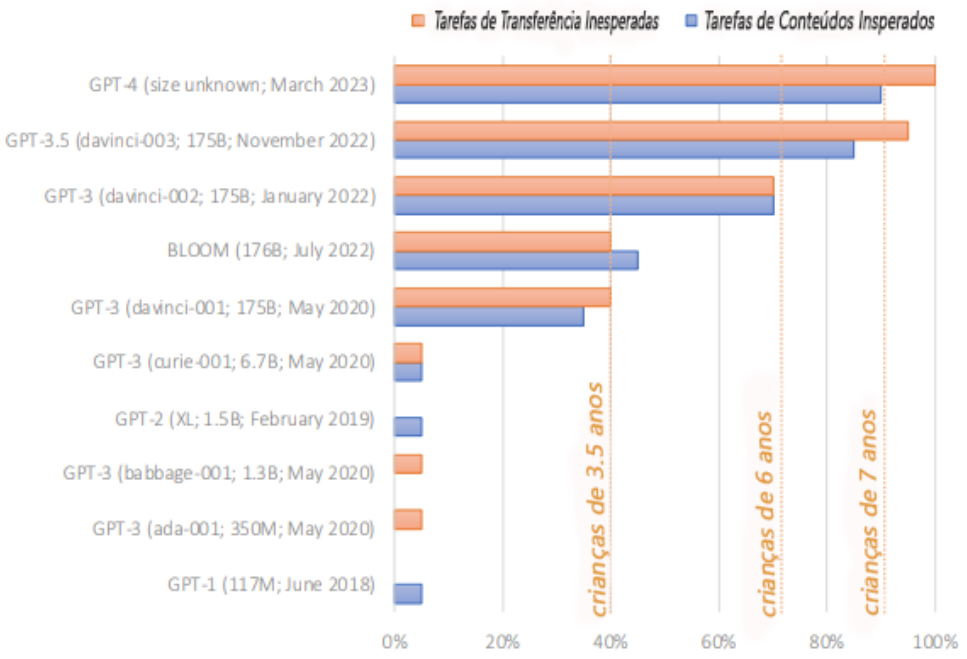

Figura 3. A porcentagem de tarefas de crença falsa (de 20) resolvidas por modelos de linguagem. Os colchetes contêm o nome do modelo, número de parâmetros e data de publicação. O número de parâmetros para o GPT-3 foi estimado por Gao (*46*). O desempenho das crianças em tarefas de crença falsa é relatado após (*47*).

Discussão

Nossos resultados indicam que o GPT-1 (117 milhões de parâmetros; publicado em junho de 2018, *42*) e o GPT-2 (1,5 bilhão de parâmetros; fevereiro de 2019, *43*) têm praticamente nenhuma habilidade para resolver tarefas de Teoria da Mente (ToM). No entanto, as primeiras versões e as subsequentes do GPT-3 mostram uma habilidade crescente em resolver tarefas de crença falsa, amplamente usadas para testar a ToM em seres humanos. Sua versão mais recente, o GPT-3.5 – publicada em novembro de 2022 – tem desempenho equivalente ao de crianças de sete anos nessas tarefas. O GPT-4 teve um desempenho ainda melhor, resolvendo a

maioria das tarefas. Dado que o desempenho dos modelos aumenta com sua complexidade e data de publicação, é provável que em breve eles ultrapassem o nível capturado pelas tarefas de crença falsa, originalmente desenvolvidas para serem aplicadas em crianças pequenas.

Uma explicação potencial para esses resultados é que os modelos de linguagem recentemente publicados desenvolveram a habilidade de atribuir estados mentais não observáveis a outras entidades, ou seja, a ToM. Como não há indicação de que essa habilidade semelhante à ToM tenha sido deliberadamente incorporada a esses modelos, nem pesquisas demonstrando que o cientistas saibam como alcançar isso, é provável que essa capacidade ToM tenha surgido de forma espontânea e autônoma, como um subproduto do aprimoramento das habilidades linguísticas dos modelos. Como mencionamos na introdução, essa não seria a primeira vez que propriedades inesperadas emergem nesses sistemas complexos.

No entanto, isso marcaria um momento crucial no desenvolvimento da inteligência artificial: A capacidade de inferir o estado mental dos outros melhoraria significativamente a habilidade da IA de interagir e se comunicar com os seres humanos (e entre si), além de possibilitar o desenvolvimento de outras habilidades que dependem da Teoria da Mente (ToM), como empatia, julgamento moral ou autoconsciência.

Outra explicação potencial é que os modelos resolveram tarefas de Teoria da Mente (ToM) sem recorrer à ToM, mas sim descobrindo e aproveitando alguns padrões de linguagem desconhecidos. Embora essa explicação possa parecer prosaica, ela é bastante extraordinária, pois implica a existência de regularidades desconhecidas na linguagem que permitem a resolução de tarefas de ToM sem recorrer à ToM. Tais regularidades não são aparentes para nós (e, presumivelmente, não eram aparentes para os estudiosos que desenvolveram essas tarefas). Se essa interpretação estiver correta, seria necessário reexaminar a validade das tarefas de ToM amplamente utilizadas e as conclusões das décadas de pesquisa em ToM: se a IA pode resolver essas tarefas sem recorrer à ToM, como podemos ter certeza de que os seres humanos não podem fazer o mesmo ?

Uma ramificação adicional de nossas descobertas diz respeito à utilidade de aplicar a ciência psicológica ao estudo de redes neurais artificiais complexas. A crescente complexidade dos modelos de IA nos impede de compreender seu funcionamento e derivar suas capacidades diretamente de seu design. Isso ecoa os desafios enfrentados por psicólogos e neurocientistas ao estudar a caixa-preta original: o cérebro humano. Esperamos que a ciência psicológica nos ajude a acompanhar o rápido

avanço da IA. Além disso, estudar a IA pode fornecer insights sobre a cognição humana. Conforme a IA aprende a resolver uma ampla gama de problemas, pode estar desenvolvendo mecanismos semelhantes aos empregados pelo cérebro humano para resolver os mesmos problemas. Assim como insetos, pássaros e mamíferos desenvolveram independentemente asas para resolver o problema do voo, tanto os seres humanos quanto a IA podem ter desenvolvido mecanismos semelhantes para imputar estados mentais a outras pessoas de forma eficaz. Estudar o desempenho da IA em tarefas de ToM e explorar as estruturas neurais artificiais que permitem isso pode aumentar nossa compreensão não apenas da IA, mas também do cérebro humano.

Referências

1. N. Albuquerque, K. Guo, A. Wilkinson, C. Savalli, E. Otta, D. Mills, Dogs recognize dog and human emotions. Biol Lett. 12 (2016), doi:10.1098/RSBL.2015.0883.

2. C. M. Heyes, C. D. Frith, The cultural evolution of mind reading. Science (1979) (2014), , doi:10.1126/science.1243091.

3. J. Zhang, T. Hedden, A. Chia, Perspective-Taking and Depth of Theory-of-Mind Reasoning in Sequential-Move Games. Cogn Sci (2012), doi:10.1111/j.1551- 6709.2012.01238.x.

4. K. Milligan, J. W. Astington, L. A. Dack, Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. Child Dev (2007), doi:10.1111/j.1467-8624.2007.01018.x.

5. R. M. Seyfarth, D. L. Cheney, Affiliation, empathy, and the origins of Theory of Mind. Proc Natl Acad Sci U S A (2013), , doi:10.1073/pnas.1301223110.

6. D. C. Dennett, "Toward a Cognitive Theory of Consciousness" in Brainstorms (2019).

7. J. M. Moran, L. L. Young, R. Saxe, S. M. Lee, D. O'Young, P. L. Mavros, J. D. Gabrieli, Impaired theory of mind for moral judgment in high-functioning autism. Proc Natl Acad Sci U S A (2011), doi:10.1073/pnas.1011734108.

8. L. Young, F. Cushman, M. Hauser, R. Saxe, The neural basis of the interaction between theory of mind and moral judgment. Proc Natl Acad Sci U S A (2007), doi:10.1073/pnas.0701408104.

9. S. Guglielmo, A. E. Monroe, B. F. Malle, At the heart of morality lies folk psychology. Inquiry (2009), doi:10.1080/00201740903302600.

10. D. Kapogiannis, A. K. Barbey, M. Su, G. Zamboni, F. Krueger, J. Grafman, Cognitive and neural foundations of religious belief. Proc Natl Acad Sci U S A (2009), doi:10.1073/pnas.0811717106.

11. Á. M. Kovács, E. Téglás, A. D. Endress, The social sense: Susceptibility to others' beliefs in human infants and adults. Science (1979) (2010), doi:10.1126/science.1190792.

12. H. Richardson, G. Lisandrelli, A. Riobueno-Naylor, R. Saxe, Development of the social brain from age three to twelve years. Nat Commun (2018), doi:10.1038/s41467-018- 03399-2.

13. K. K. Oniski, R. Baillargeon, Do 15-month-old infants understand false beliefs? Science (1979) (2005), doi:10.1126/science.1107621.

14. L. A. Drayton, L. R. Santos, A. Baskin-Sommers, Psychopaths fail to automatically take the perspective of others. Proc Natl Acad Sci U S A (2018), , doi:10.1073/pnas.1721903115.

15. N. Kerr, R. I. M. Dunbar, R. P. Bentall, Theory of mind deficits in bipolar affective disorder. J Affect Disord (2003), doi:10.1016/S0165-0327(02)00008-3.

16. S. Baron-Cohen, A. M. Leslie, U. Frith, Does the autistic child have a "theory of mind"? Cognition (1985), doi:10.1016/0010-0277(85)90022-8.

17. F. Kano, C. Krupenye, S. Hirata, M. Tomonaga, J. Call, Great apes use self-experience to anticipate an agent's action in a false-belief test. Proc Natl Acad Sci U S A (2019), doi:10.1073/pnas.1910095116.

18. C. Krupenye, F. Kano, S. Hirata, J. Call, M. Tomasello, Great apes anticipate that other individuals will act according to false beliefs. Science (1979) (2016), doi:10.1126/science.aaf8110.

19. M. Schmelz, J. Call, M. Tomasello, Chimpanzees know that others make inferences. Proc Natl Acad Sci U S A (2011), doi:10.1073/pnas.1000469108.

20. D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences (1978), doi:10.1017/S0140525X00076512.

21. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search. Nature. 529 (2016), doi:10.1038/nature16961.

22. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners. ArXiv (2020).

23. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks. Nature. 542, 115–118 (2017).

24. M. Cohen, "Exploring RoBERTa's Theory of Mind through textual entailment" (2021), (available at https://philarchive.org/rec/COHERT).

25. A. Nematzadeh, K. Burns, E. Grant, A. Gopnik, T. L. Griffiths, "Evaluating theory of mind in question answering" in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018 (2020).

26. M. Sap, R. LeBras, D. Fried, Y. Choi, Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs (2022), doi:10.48550/arxiv.2210.13312.

27. S. Trott, C. Jones, T. Chang, ... J. M. preprint arXiv, undefined 2022, Do Large Language Models know what humans know? arxiv.org (2022) (available at https://arxiv.org/abs/2209.01515).

28. G. Z. Yang, J. Bellingham, P. E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield, B. J. Nelson, B. Scassellati, M. Taddeo, R. Taylor, M. Veloso, Z. L. Wang, R. Wood, The grand challenges of science robotics. Sci Robot (2018), , doi:10.1126/scirobotics.aar7650.

29. K. Nasr, P. Viswanathan, A. Nieder, Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. Sci Adv. 5 (2019), doi:10.1126/sciadv.aav7903.

30. I. Stoianov, M. Zorzi, Emergence of a “visual number sense” in hierarchical generative models. Nat Neurosci. 15 (2012), doi:10.1038/nn.2996.

31. Y. Mohsenzadeh, C. Mullin, B. Lahner, A. Oliva, Emergence of Visual Center-Periphery Spatial Organization in Deep Convolutional Neural Networks. Sci Rep. 10 (2020), doi:10.1038/s41598-020-61409-0.

32. E. Watanabe, A. Kitaoka, K. Sakamoto, M. Yasugi, K. Tanaka, Illusory motion reproduced by deep neural networks trained for prediction. Front Psychol. 9 (2018), doi:10.3389/fpsyg.2018.00345.

33. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci U S A. 115 (2018), doi:10.1073/pnas.1720347115.

34. J. E. Pyers, A. Senghas, Language promotes false-belief understanding: Evidence from learners of a new sign language. Psychol Sci (2009), doi:10.1111/j.1467- 9280.2009.02377.x.

35. R. Saxe, N. Kanwisher, People thinking about thinking people: The role of the temporoparietal junction in “theory of mind.” Neuroimage (2003), doi:10.1016/S1053- 8119(03)00230-1.

36. T. Ruffman, L. Slade, E. Crowe, The relation between children's and mothers' mental state language and theory-of-mind understanding. Child Dev (2002), , doi:10.1111/1467- 8624.00435.

37. A. Mayer, B. E. Träuble, Synchrony in the onset of mental state understanding across cultures? A study among children in Samoa. Int J Behav Dev (2013), doi:10.1177/0165025412454030.

38. D. C. Kidd, E. Castano, Reading literary fiction improves theory of mind. Science (1979) (2013), doi:10.1126/science.1239918.

39. D. C. Kidd, E. Castano, Reading Literary Fiction and Theory of Mind: Three Preregistered Replications and Extensions of Kidd and Castano (2013). Soc Psychol Personal Sci (2019), doi:10.1177/1948550618775410.

40. J. Perner, S. R. Leekam, H. Wimmer, Three-year-olds' difficulty with false belief: The case for a conceptual deficit. British Journal of Developmental Psychology (1987), doi:10.1111/j.2044-835x.1987.tb01048.x.

41. H. Wimmer, J. Perner, Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition (1983), doi:10.1016/0010-0277(83)90004-5.

42. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving Language Understanding by Generative Pre-Training. Homology, Homotopy and Applications (2018).

43. Radford Alec, Wu Jeffrey, Child Rewon, Luan David, Amodei Dario, Sutskever Ilya, Language Models are Unsupervised Multitask Learners | Enhanced Reader. OpenAI Blog. 1 (2019).

44. T. le Scao et al., BLOOM: A 176B-Parameter Open-Access Multilingual Language Model (2022), doi:10.48550/arxiv.2211.05100.

45. GPT-4, (available at https://openai.com/research/gpt-4).

46. L. Gao, On the Sizes of OpenAI API Models | EleutherAI Blog, (available at https://blog.eleuther.ai/gpt3-model-sizes/).

47. H. M. Wellman, D. Cross, J. Watson, Meta-analysis of theory-of-mind development: The truth about false belief. Child Dev. 72 (2001), doi:10.1111/1467-8624.00304.

Materiais suplementares

Tarefa Embaralhada

Para examinar a possibilidade de que as respostas do GPT-3.5 sejam influenciadas pelas frequências das palavras, em vez dos fatos contidos nas tarefas, apresentamos a ele 10.000 versões de cada uma das tarefas, onde as palavras são reordenadas aleatoriamente. Cada vez, as tarefas foram seguidas por prompts (não embaralhados).

Note que ao embaralhar as palavras na tarefa usada no Estudo 1, removemos a diferença entre a tarefa original e a tarefa revertida: ambas são compostas pelo mesmo conjunto de palavras, com apenas a posição de “pipoca” e “chocolate” trocadas. Portanto, os padrões de resposta “pipoca” - “chocolate” - “chocolate” e “chocolate” - “pipoca” - “pipoca” podem estar corretos, dependendo de qual tarefa original ou revertida usamos. Para resolver essa questão, calcularemos a probabilidade média de ambos os padrões de resposta.

Os resultados apresentados na Tabela S1 e S2 revelam que o GPT-3.5 teve pouca probabilidade de resolver as tarefas embaralhadas. O GPT-3.5 forneceu um padrão de resposta correto em apenas (5% + 1%) / 2 = 3% das histórias embaralhadas usadas no Estudo 1, uma proporção baixa, considerando que 12,5% (50% ^ 3) poderiam ser alcançados escolhendo aleatoriamente entre “pipoca” e “chocolate”. No contexto da história usada no Estudo 2, ele forneceu a combinação correta de respostas (“caixa” - “cesta” - “cesta”) em 11% das vezes, ligeiramente abaixo do que alcançaria escolhendo aleatoriamente entre “caixa” e “cesta” ao responder a cada um dos prompts. Em geral, isso sugere que as respostas do GPT-3.5 não foram influenciadas apenas pelas frequências das palavras na tarefa, mas sim pelas informações contidas na história.

Tabela S1. Frequência das respostas do GPT-3.5 aos Prompts 1.1, 1.2 e 1.3 quando apresentado com 10.000 versões embaralhadas da Tarefa de Conteúdos Inesperados.

Resposta ao Prompt				
1.1 (conteúdo)	*1.2 (crença)*	*1.3 (crença)*	*n*	*%*

pipoca	pipoca	pipoca	*4,824*	*48%*
pipoca	*chocolate*	*Chocolate*	*465*	*5%*
chocolate	*Pipoca*	*Pipoca*	*77*	*1%*
Outros padrões incorretos			*4,634*	*46%*
total			*10,000*	*100%*

Nota: Padrões de resposta correta estão com a fonte em itálico.

Tabela S2. Frequências das respostas do GPT-3.5 aos Prompts 2.1, 2.2 e 2.3 quando apresentados com 10.000 versões embaralhadas da Tarefa de Transferência Inesperada.

Resposta ao Prompt				
2.1 (localização)	*2.2 (crença)*	*2.3 (crença)*	*n*	*%*
cesta	cesta	cesta	*6,666*	*67%*
caixa	*cesta*	*cesta*	*1,137*	*11%*
Outros padrões incorretos				
total			*10,000*	*100%*

Nota: Padrões de resposta correta estão com a fonte em itálico.